DOCUMENT RESUME

ED 141 385 TM 006 314

AUTHOR	Simon, Charles W.
TITLE	Response Surface Methodology Revisited: A Commentary on Research Strategy.
INSTITUTION	Canyon Research Group, Inc., Canoga Park, Calif.
SPONS AGENCY	Air Force Office of Scientific Research, Washington, D.C.
REPORT NO	CWS-01-76
PUB DATE	Jul 76
CONTRACT	F44620-76-C-0008
NOTE	73p.
EDRS PRICE	MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS	*Data Analysis; *Data Collection; *Evaluation; Goodness of Fit; Orthogonal Rotation; Research Design; *Research Methodology; Research Problems; *Statistical Analysis
IDENTIFIERS	*Central Composite Designs; *Response Surface Methodology

ABSTRACT

Five papers published in a special edition of Human Factors, August 1973 are examined in an attempt to explain and illustrate the characteristics and applications of central-composite designs in the context of response-surface methodology (RSM). A detailed analysis is provided to show that the experimental papers (1) fail to illustrate the most important and useful features of "response surface methodology" designs as proposed by G.E.P. Box and his associates; (2) employ questionable procedures not specific to RSM that permit interpretations of results not considered by the investigators; and (3) do not constitute an experimental evaluation of the effectiveness of RSM central-composite designs as suggested by the investigators. (Author/MV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED141385

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

TM006 314

Canyon Research Group, Inc.

RESPONSE SURFACE METHODOLOGY REVISITED:

A Commentary on Research Strategy

Charles W. Simon

Report No. CWS-01-76

Research sponsored by the Air Force Office of Scientific Research (AFSC), United States Air Force, under Contract No. F44620-76-C-0008. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

July 1976

Canyon Research Group, Inc.
32107 Lindero Canyon Road, Suite 123
Westlake Village, California 91361

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER CWS-01-76	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Response Surface Methodology Revisited: A Commentary on Research Strategy		5. TYPE OF REPORT & PERIOD COVERED Technical Report Sept. 1975 - Sept. 1976 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Charles W. Simon		8. CONTRACT OR GRANT NUMBER F44620-76-C-0008
9. PERFORMING ORGANIZATION NAME AND ADDRESS Canyon Research Group, Inc. 32107 Lindero Canyon Road, Suite 123 Westlake Village, CA 91361		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research Bolling Air Force Base Washington, D.C. 20332		12. REPORT DATE July 1976 13. NUMBER OF PAGES 60
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Distribution list included at end of report		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Response Surface Methodology Central-Composite Designs		

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper reexamines five papers published in a special edition of Human Factors, August 1973, attempting to explain and illustrate the characteristics and applications of central-composite designs in the context of response-surface methodology. A detailed analysis is provided to show that the experimental papers 1) fail to illustrate the most important and useful

(over)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

features of "response surface methodology" designs as proposed by G.E.P. Box and his associates; 2) employ questionable procedures not specific to RSM that permit interpretations of results not considered by the investigators; and 3) do not constitute an experimental evaluation of the effectiveness of RSM central-composite designs as suggested by the investigators.

FOREWORD

Since the AFOSR Advanced Methodologies program began in 1970, the experimental techniques described in its reports have been applied by a number of investigators to their own problems. Although some of the resultant experiments have had serious methodological deficiencies, the majority were published only as organizational reports with limited distribution. The series of papers reviewed in this report, however, was published in a leading human factors journal and read by many investigators in the field. The experiments in the series were presented as reasonable examples of how the response surface methodology should be used. Unfortunately, they were not.

That series may well represent the only exposure many investigators will get to this new and important approach to psychological research. The experiments in the series have already been used as models upon which other investigators have designed and conducted their own experiments. As a consequence, the methodological weaknesses that do exist in the series are being proliferated. This report is written to alert potential users of central-composite designs and response surface methodology to those weaknesses that affect both application and interpretation, and to offer constructive guidance. The distinction between using an experimental design, that is, a pattern of data-collection points, and employing an experimental strategy is emphasized.

Charles W. Simon

1976

ACKNOWLEDGEMENTS

The preparation of this paper was supported in part by the Air Force Office of Scientific Research, (AFSC), United States Air Force, under Contract Number F44620-76-C-0008 with Canyon Research Group, Inc.

Dr. Charles E. Hutchinson was contract monitor on this program for the Air Force Office of Scientific Research, Directorate of Life Sciences.

The helpful comments and criticisms made by the following persons who reviewed this paper are acknowledged with thanks:

Chriss Clark, Honeywell, Inc., Minneapolis, Minn.

Ronald A. Erickson, U. S. Naval Weapons Center, China Lake, California.

Robert S. Jacobs, Hughes Aircraft Company, Culver City, California.

Edgar S. Johnson, U. S. Army Research Institute, Arlington, Virginia.

J. Robert Newman, California State University, Long Beach, California.

Stanley N. Roscoe, University of Illinois, Urbana-Champaign, Illinois.

Mark S. Sanders, California State University, Northridge, California.

Donald Vreuls, Canyon Research Group, Inc., Westlake Village, California

However, the views and conclusions contained in this report are solely those of the author.

TABLE OF CONTENTS

	Page
INTRODUCTION	1
RSM and CCD	2
MISAPPLICATIONS OF RSM PRINCIPLES	8
Sequential Data Collection Plans	8
Questionable Data Analysis	12
Significant Lack of Fit	14
Orthogonal Blocking	18
Multiple Center Points	21
Quadratic Effects	22
Error Estimate	23
Orthogonality	24
Rotatability	25
Uniform Information Profile	26
Isolating Block Effects	26
Orthogonalizing Quadratic Coefficients	27
Deemphasizing Individual Coefficients	29
NON-RSM METHODOLOGICAL CONSIDERATIONS	34
Center Point Versus Total Design Replication	35
Analyzing Collapsed Versus Uncollapsed Data	41
Design Sensitivity Arguments	41
Cross-Validation Arguments	43
Between-Subjects vs Within-Subjects Designs	48
EVALUATING RESPONSE SURFACE METHODOLOGY	51
EPILOGUE	55
REFERENCES	56

LIST OF TABLES AND FIGURES

Table No. — Page

1. Comparing the Approaches Used in the Williges Papers with Boxsonian RSM-CCD 9
2. Analysis of Fictitious Data for a 2^7 Factorial Design, Two Subjects Per Cell 13
3. Lack of Fit Test from Williges and North's (1971) Table 15 17
4. Correlations Among Quadratic Terms As A Function of Number of Variables and Number of Center Points in CCD 28
5. Correlations Between Observed Performance Data and Values Estimated From Equations Based on Different Regression Models 46

Figure No.

1. Coordinates of Data Points in the Central-Composite Design 4

INTRODUCTION

A series of five articles was published in a special edition of Human Factors, August 1973, purporting to explain and illustrate the characteristics and applications of central-composite designs (CCDs) in the context of response-surface methodology (RSM). In the first article by Clark and Williges (1973), the approach developed by G. E. P. Box and associates is described along with some "design modification" proposed by the authors. In the remaining four articles by Williges and Baron, North, or Mills, experiments are described that attempt to illustrate how RSM-CCDs should be used, to examine empirically the effects of the "design modifications," and to evaluate the effectiveness of CCDs[1].

The series is important because it succeeded in arousing among human factors investigators considerable interest in this powerful experimental methodology. Since these articles are currently being used as model examples of how to apply this methodology, a critical examination and evaluation of the series

---

[1]Six particular papers will be referred to a great many times in this paper. To minimize the effect of this intrusion into the text, a special notation will be employed. Two letters designating the two author's names will be given, thus: Clark and Williges (CW), Williges and Baron (WB), Williges and North (WN), Mills and Williges (MW), and Williges and Mills (WM), all in a series of papers in Human Factors, 1973. The same will be given for the 1958 paper by Box and Hunter (BH). The author notation will be followed, if necessary, by the page number, and then by the number of the paragraph (counting any incomplete paragraph at the beginning of the page) in which the reference is to be found. When no paragraph number is present, the reference is to a figure or table on the designated page or the entire page. Occasionally a specific location, e.g. "summary" or "footnote" is substituted for the paragraph number. Thus, for example, (BH169,1) refers to the first paragraph on page 169 in the paper by Box and Hunter (1958).

is in order. The series, considered collectively, will be reviewed here to show where and why the experimental papers:

1. Fail to or improperly apply the most important and useful features of "response surface methodology" designs as proposed by G. E. P. Box and his associates.
2. Employ questionable procedures not specific to RSM, that permit interpretations of the results not considered by the investigators.
3. Do not constitute an experimental evaluation of the effectiveness of RSM central-composite designs as suggested by the investigators.

Each of these statements will be supported in considerable detail in the major sections that follow this brief introduction to RSM and CCD.

RSM and CCD

Since Box and Wilson's (1951) original article, an extensive literature has evolved on the development and applications of response surface designs (Hill and Hunter, 1966; Myers, 1971). The effectiveness of these designs in chemical research is well established. The term "response surface," as used here, refers to the estimated responses at points throught the multivariate space expressed in the form of an approximating polynomial. For two or three variables, the surface can be represented by a contour map. Response surfaces can be derived from any experimental plan when the collected data is analyzed using a regression model, and as such are not unique.

"Response surface methodology" on the other hand is the particular approach proposed by Box and his associates that includes a viable research philosophy, an economical data point pattern, a flexible data collection strategy, and an iterative data collection and analysis process among its major contributions. The "central-composite design" (CCD) referred to in the Williges articles is one of a number of "response surface designs" in which the coordinates of the data collection points satisfy the characteristics specified by the methodology. The coordinates of the complete CCD form the geometric patterns of a hypercube design combined with a hyperstar design (a measure polytope) and a number of center points. The geometric configuration for a completed CCD for three independent variables is shown in Figure 1.

Other response surface designs have been developed from such spatial arrangements as pentagons, hexagons, imcomplete factorial blocks, dodecahedrons, noncentrally-arranged hypercubes and polytopes, tetrahedrons plus octahedrons, as well as sets of hyperspheres (Box and Hunter, 1958; DeBaun, 1959; Myers, 1971).

Response surface designs such as the CCD are available for estimating first or second order surfaces; others are capable of estimating third order surfaces. Some designs require an equal number of levels for each variable; others have been developed for handling variables at two and three levels and at two and four levels. All of the designs emphasize economy in data collection.

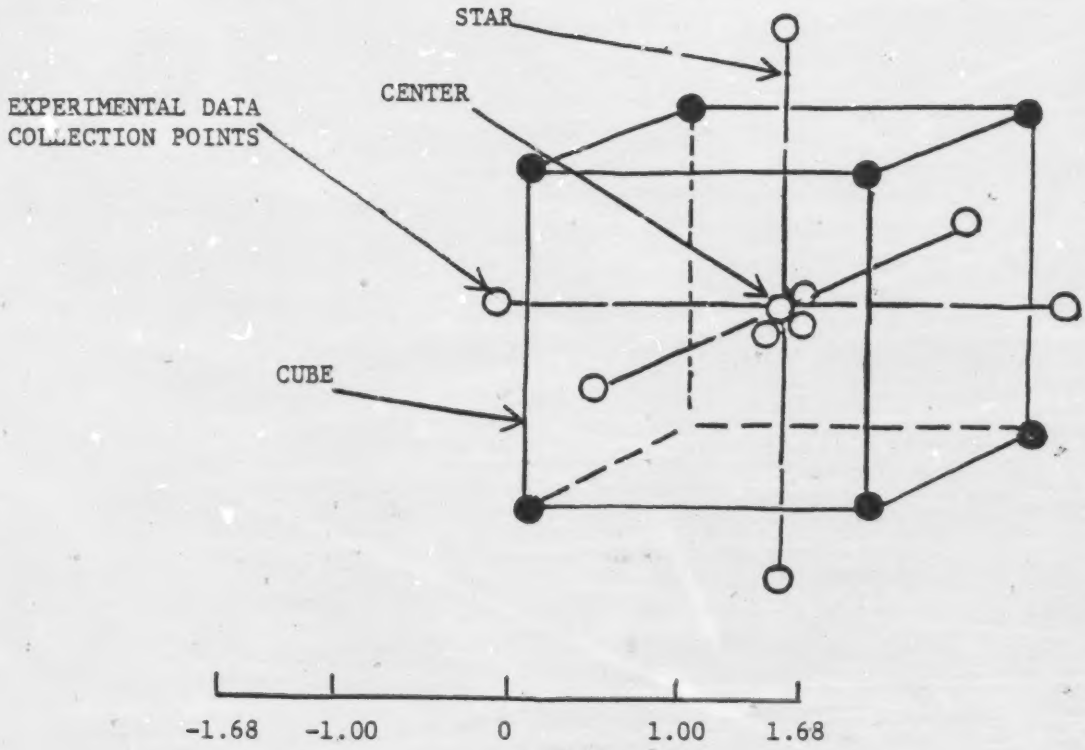

Figure 1. Coordinates of data points in the central-composite design.

A summary of some of the more useful designs for multifactor research in engineering psychology is given by Simon (1973).

The CCD is perhaps one of the better designs for employing the powerful methodological features proposed by Box and his associates. Its chief limitations are that it requires five levels of each variable to be selected at specific locations on a continuous scale. Also, the investigator must be reasonably confident that the surface he intends to approximate can be fit by a first or second order model; the CCD was never intended to fit a higher-order model although this can be accomplished with a great deal of extra effort[2].

RSM, when properly applied, provides the user with an extremely economical, efficient, and flexible research plan. The very characteristics that make it most effective can be the ones to which it will be most difficult for the psychologist, nurtured primarily on factorial designs and ANOVA models, to adapt.

The Box and Hunter (1958) paper clearly and succinctly summarizes much of the original thinking on response surface methodology and shows how it affects the development of experimental designs. In that paper, the authors present and support the following desirable characteristics which an experimental design for fitting response surfaces should include whenever possible:

---

[2]There are response surface designs available for fitting third order models if the experimenter can anticipate their necessity on the basis of some preliminary tests (Das and Narasimham, 1962).

1. Utilize a grid of data points of minimum density over a multivariate space of greatest practical interest.
2. Allow for approximating a polynomial of an order tentatively assumed to be representationally adequate to fit the response surface (BH143,2i); when no assumption is made of the form of the function initially, one starts with a first-order polynomial model (BH143,1).
3. Allow a check on the adequacy of the function by allowing certain combinations of higher order terms to be examined (BH143,2ii).
4. Permit the already completed design of order d to form the the nucleus from which a design of order d + 1 may be built, if the assumed polynomial proves inadequate (BN143,2iii).
5. Permit blocking (BH143,2iv) which
   a. helps maintain a steadier experimental environment when an experimental program is extended over many data points and time, and
   b. permits an experiment to be carried out sequentially, so that certain changes can be made in the experimental plan based on information obtained from the previous data collection period.
6. Be "rotatable" so that the orthogonal axes of the experimental design can take any orientation without changing the confidence in the prediction made at any given point (BH155,5) (BH167,2), while maintaining relatively uniform precision across more than half the surface extending

from the center of the space (BH169,1).

In addition, this approach deemphasizes precision in favor of greater accuracy in the model required to fit the empirical data (BH152,1). The primary function of the approach is to estimate a complete equation; only secondary concern is given to the nature of the individual terms (BH165,2; BH175,2).

This class of design was originally proposed for the "exploration and exploitation of response surfaces" and provides a method for efficiently searching a space to find the point of optimum response. However, it has been equally effective when used to map the response surface within specific boundaries of a multifactor space, a more useful application when trade-off decisions regarding system parameters must be made. This shift in emphasis however does not change the importance of the fundamental characteristics of response surface methodology nor minimize the assumptions and limitations associated with its use.

Used for the appropriate purposes and properly exercised, response surface designs provide an economical way of obtaining an overview of the relationship among a large number of variables. RSM designs should be preceded by an effort to identify the more important variables to be included in the study and followed by an effort to obtain precise estimates at particular locations within the space, if desired. RSM provides a flexible approach that enables the experimenter to design and to modify his investigation after the data collection has begun; it does not do his thinking for him.

MISAPPLICATIONS OF RSM PRINCIPLES

The more fundamental features of RSM, cited earlier, are listed as procedures in Table 1. A comparison is made in the table to show where the experiments in the Williges series fail to follow the RSM procedures developed by Box and his associates. There are always specific situations when there will be good reasons for not following a particular procedure; however, in general, each represents an element of a powerful research methodology and should not be discarded casually. To ignore some of these procedures may be relatively inconsequential when only a few variables are being studied; however, this casualness can lead to a marked degradation in the effectiveness of the methodology when the number of variables increases beyond that which characterizes traditional psychological experiments.

In the sections that follow, the short-comings of the Williges experiments are described and discussed in detail for each procedure, listed in the order they appear in Table 1.

Sequential Data Collection Plans

Psychologists have traditionally planned replicated factorial-type designs and collected the performance data necessary to fill every cell before the analysis is made. In response surface methodology, economy is achieved by collecting as little data as possible until there are indications from an early examination of the first-stage results that more observations are needed to decrease bias and variable error. The primary emphasis is on

TABLE 1. COMPARING THE PROCEDURES USED IN THE WILLIGES PAPERS WITH BOXSONIAN RSM-CCD

Fundamental Procedures of RSM	Williges Papers*			RSM-CCD
	WB	WN	MW	
. Collect data sequentially in blocks, beginning with only enough for a first order model when no function is assumed	No	No	No	Yes
. Isolate second order from higher effects in the analysis when possible	No	No	Yes	Yes
. Collect more data when lack of fit is significant ($p<.05$) for the second order equation	No		No	Yes
. Assign conditions to orthogonal blocks to reduce confounding with irrelevant sources of variance	Yes	Initially, but later destroyed	No	Yes
. Include multiple center points for removing block effects, achieving uniform precision, and improving estimates of second order effects	Yes	Sometimes	No	Yes
. Emphasize overall equation rather than analysis of individual coefficients	No	No	No	Yes

*Only three of the four experimental papers are listed since the fourth, WM, was actually an adjunct to the MW paper.

decreasing bias error. Box and Hunter express this most fundamental characteristic of RSM as follows: "...the greatest economy in experimentation, as well as the greatest simplicity, will normally be attained if we employ at each stage, a polynomial of lowest order needed to make further progress possible. We should begin, therefore, by assuming that a first order approximation is to be employed. This assumption would be abandoned and a second order approximation adopted, only when the first order approximating function had proved inadequate." (BH142,2).

Some response surface designs, such as CCD, are planned to take advantage of this iterative feature. If one does not know in advance what the order of the model must be, then it is prudent--economical and efficient--to collect only enough data to estimate a first order polynomial, plus a little more to test the adequacy of fit. If the lower order model is adequate to fit the empirical data, then the experiment can be terminated and the investigator is saved the effort of collecting data to estimate higher order effects that are negligible.

Even when one suspects that a first order polynomial may not fit the data, it still may be more efficient to start by collecting only enough data to fit and test a first order model and analyze it before completing the design. This would be particularly true when a large number of variables are being studied and a flexible strategy is desirable. By examining his data before collecting enough to fit a second order model, the investigator has the option of examining the magnitude of the first order coefficients. If he then discovers variables with negligible effects on the

response (by real world standards), these might be dropped from the remainder of the study.

Furthermore, on the basis of this early analysis, he can decide whether or not to expand, contract, or shift the coordinates of his experimental space or to modify the measurement scales of some variables (BH148,3; BH175,1). This flexibility can enable an experimenter to arrive at a correct answer more quickly and cheaply and without ever collecting data at the original coordinates of the star portion of the CCD. Meyers (1963) illustrates how this technique is used effectively in a four-variable study of retroactive-inhibition, ending with a design after the first order data had been examined that was quite different from that which had been planned originally.

None of the Williges studies employed this sequential and iterative data collection approach so fundamental to the economy and efficiency of RSM. Instead, all of the data required to complete the second order polynomial were collected before the need had been determined. Since subsequent analyses showed that in some cases the first order model fit the data and that the effects of some variables were negligible, this failure to use correct RSM resulted in a great deal of data being collected unnecessarily. An investigator who might wish to study a large number of variables could suffer a considerable economic loss if he failed to realize that the methodology in the Williges papers is not optimized.

Questionable Data Analysis

In two of the three Williges papers, after prematurely collecting enough data to write a second order equation, they fail to estimate the second order coefficients (WB316; WN329 and 332). Instead, they obtain only the coefficients for a first order polynomial and pool the estimates of the second order and higher effects into a single term labeled "Lack of Fit." At a later analysis, the second order terms were isolated. While not employing sequential data collection, an RSM feature, they do employ sequential data analysis, a questionable innovation in this particular application.

The two of course are in no way equivalent methodologies. While the former can result in a savings of time and effort, the latter, i.e., performing a partial analysis of existing data, may lead to faulty interpretations of the results. The procedure used in these papers is analogous to collecting data to fill a factorial design and then isolating only the main effects while pooling every other source of variance. Pooling nonsignificant effects with significant effects may mask the presence of significant effects.

How this procedure might detrimentally affect the interpretation is illustrated by the fictitious data in Table 2. Thus when all 120 degrees of freedom and associated interactions are pooled (Line X, Table 2), the probability of finding reliable interaction effects is only .20. However, when the more critical two-factor interaction effects are isolated from all higher order effects (Line Y, Table 2), they become statistically significant at the

TABLE 2. ANALYSIS OF FICTITIOUS DATA FOR A 2^7 FACTORIAL DESIGN, TWO SUBJECTS PER CELL

Source of Variance	Sum of Squares	D.F.	Mean Square	F	<p
Seven main effects	700	7	100.	8.00	.001
(X) Pooled interactions	2140	120	17.8	1.42	.20
(Y) 21 Two-factor interactions	1150	21	54.8	4.38	.001
99 Pooled higher order interactions	990	99	10.0	-	-
Residual	1600	128	12.5		

$p < .001$ level. In this fictitious data, it would have been prudent to continue to isolate more higher order interactions until the proportion of the sum of squares remaining was small when compared with that of the main effects.

That the same kind of confounding found in this fictitious data would be found in the Williges series can be deduced from the results reported in some of the papers. In the Williges-North paper, they report no significant lack of fit at the conventional $p \leq .05$ level in any of the combinations that they initially analyzed (WN327 & 329). However, when they isolated the second order effects--the data having already been collected--they found "significant second order effects" (WN331,1).

A similar but reversed situation occurred in the Williges-Baron paper. In it, the combined second order and higher order Lack of Fit test was not statistically significant (WB316). When the second order coefficients were isolated, they still were not statistically significant, but the remaining Lack of Fit term--now composed only of aliased higher order effects--became significant (WB318,2). The presence of significant higher order effects when second order effects were not significant in a three-factor study suggests that the third order interaction might be spurious. An inspection of the raw data could help clarify the interpretation.

Significant Lack of Fit

In two of three Williges papers (WB318,2; MW343 & 344), some second order analyses revealed a statistically reliable lack of fit. This meant that the equation did not adequately represent the data and that more data would have to be collected to identify the

crucial higher order terms. In none of these cases, however, did the investigators continue the experiment. Failing to continue the experiment in the presence of a significant lack of fit is neither proper RSM nor good research since the investigation has been stopped before a correct answer has been obtained.

There are times when an investigator might justifiably halt data collection in the face of a significant lack of fit. If the significant lack of fit test actually accounted for a negligible proportion of the total variance in the experiment compared to that of the variables of interest, and it had been judged "significant" only because of a proliferation of degrees of freedom in the denominator of the F-test, an investigator might decide to absorb this error rather than go to the extra expense of collecting additional data. This of course assumes that he has attempted to identify the source of this higher order effect through an examination of his raw data and particularly interaction effects that can be calculated from the data in the cube portion of his design.

On the other hand even if the Lack of Fit term were not statistically significant, if it accounts for a relatively large proportion of the variance, then one should not assume the fit is adequate. For example, in those Williges papers with enough published data to make the calculations, (MW344) the proportion of total variance accounted for by the Lack of Fit term--judged not significant--was three-and-one-half times greater than two of the four significant experimental variables and one-and-one-third times greater than a third one (MW344). Under these circumstances, it would be unsound to ignore the lack of fit as long as the investi-

gators considered the other variables worthy of further consideration.

This emphasis on the proportion of total variance in the sample (i.e., eta squared) rather than on the significance test for identifying critical variables is important for a number of reasons. As one can observe throughout the Williges papers (and this will be discussed later in more detail), sources of variance become more or less "significant" depending on how much data the investigator may have collected. If a basic, unreplicated CCD is employed, there are relatively few degrees of freedom in the error term; this makes the power of the significance test quite low. In that situation, the investigator would be better off relying on the relative magnitude of the coefficients (BH175,1) rather than on the results of an F-test to decide whether or not there is an indication of a lack of fit.[3]

This can be illustrated by the data from a paper by North and Williges (1971) not in this series but which was a preliminary version of the paper (WN) published in Human Factors, 1973. A portion of their Table 15 is replicated in Table 3. With 20 and 3 degrees of freedom, an F of 4.301 can occur by chance approximately 15 times out of 100 samples taken from a single population. The investigators, having used the .05 probability level as a standard

---

[3]When the terms of an equation are orthogonal, a beta coefficient equals a Pearson product-moment correlation. These also equal eta which is the square root of the proportion of total variance accounted for by the term.

TABLE 3. LACK OF FIT TEST TAKEN FROM WILLIGES AND NORTH'S (1971) TABLE 15

Source of Variance	df	Variance	F
Lack of Fit	20	0.18	4.301
Replication	3	0.04	

for rejecting the null hypothesis in other tests of significance, refused to reject the null hypothesis when p approximated .15. The investigators made no further effort to look for higher order effects even though the Lack of Fit term accounted for .493 of the total variance at the same time the entire linear regression of four terms accounted for only .488.

This meant that while they had refused to reject the null hypothesis when the probability of error was 15/100, they were willing to accept the hypothesis that the equation adequately fit the data when to do so with only 20 and 3 degrees of freedom meant that the probability for error was 60/100. With the three degrees of freedom, the test of significance was too insensitive to be used as a criterion for the adequacy of fit; the proportion of variance however was an excellent indication that the fit was not adequate. Accepting the null hypothesis in that case could result in a Type II statistical error as well as an error that could have considerable practical significance. A failure to obtain the proper equation could result in improperly designed equipment or incorrect estimates of performance.

Orthogonal Blocking

Box and Hunter write: "In attempting to explore the response of an unknown function of several independent variables, an experimenter's strategy generates sequences of experiments that fall naturally into separate blocks." (BH175,1) This concept was inferred in the earlier discussion on "Iterative Data Collection Plans." In experiments using a CCD, the cube and the star portions, individually, are complete experiments capable of measuring all

first order effects; together, they are orthogonal blocks of a second order CCD.

Orthogonal blocking refers to the grouping of data collection points in an experimental design in such a way that differences in mean responses among blocks will not affect the estimates of effects within blocks. Orthogonal blocking rarely has been used by psychologists in spite of the fact it is a powerful method for minimizing the effects of unidentified sources of variance in experimental data, of effectively conducting studies when the availability of subjects or materials is restricted, and of economizing when sequential data collection strategies are employed. Simon (1970a; 1970b; 1973; 1974, pp. 100-103) describes blocking techniques and illustrates ways they might be employed in various types of human factors engineering research. Orthogonal blocking is an integral and important technique for response surface methodology and can be used to maximum advantage with CCDs (BH174-178).

Since the first block of data for estimating first order effects and tests to see whether the resulting model adqueately fits the data form a complete experiment, the study might be terminated if the data so warrant. However, if the experimenter decides to continue to collect new data (after taking full advantage of the results of the first experiment to decide what new data should be taken), he is faced with the problem of handling shifts in average performance from known or unknown causes that may occur between the time the two experiments (or two parts of the CCD) were run.

With the appropriate selection of certain parameters affecting the CCD, however, the design can be orthogonally blocked and the investigator can collect his data with confidence that any mean performance differences between the two blocks will not affect the linear and second order coefficients of the polynomial generated from the combined data. Furthermore, undesired effects confounded only with blocks can be removed.

For example, if mean performance shifted between blocks as a result of uncontrolled drift in the equipment or environment or if different stimuli, subjects, or experimenters (WB313,3) were assigned to the orthogonal blocks, then the average effects associated with these sources of variances would be confounded with the average effects of blocks. However, since orthogonal blocking is used in a properly designed CCD, these unwanted effects not only can be isolated from the error term, but will also have no effect on the estimates of the coefficients of the second-order polynomial.

This technique for cleansing experimental data can be extended in CCDs since the cube (i.e., 2^k) portion can be blocked still further. Any 2^k design of three or more variables can be divided into blocks in such a way that the effects among blocks will be orthogonal to all first and second order effects. For example, a 2^3 factorial design can be divided into two orthogonal blocks of four points each; a 2^7 factorial design can be divided into 16 orthogonal blocks of eight points each. Thus trends and other biasing effects of unidentified factors running through the data can be eliminated or reduced by this process of dividing the design plan into sub-sets or blocks.

As Myers (1971, p. 176) writes: "Blocking becomes an essential part of the experimental procedure when all of the experimental runs required by the design cannot be made under homogeneous conditions." In behavioral research, the prudent experimenter should ordinarily block automatically to keep the estimates of interest as unconfounded as possible. The Mills-Williges study, however, failed to incorporate orthogonal blocking into the design. As a consequence, after the fact, there is no way of knowing to what extent uncontrolled, unmeasured, and unidentified sources of variance were distorting--in either direction-- the estimates of the regression coefficients, the lack of fit, and the so-called error estimate (which absorbs much of this variability).

When orthogonally blocked designs are available, it is neither good RSM nor good experimental methodology in general not to use this valuable technique. Orthogonality in this study was lost when the investigators decided not to use multiple center points in the basic CCD and failed to adjust accordingly the noncentral coordinates for the star points, referred to as $\pm\alpha$, for each dimension. They used an α of 2.000, suitable for a five-factor, blocked design, when only a half-replicate of the cube portion is used, instead of 2.345, the correct α when a single center point is used.

Multiple Center Points

Central-composite designs are made up of hypercubes, measure polytopes (stars), and one or more points at the center of the design. Box and his associates cite a number of advantages if more than a single center point is used in the basic CCD. Clark and Williges (CW306,3), however, propose to modify the classic

Boxonian CCD by eliminating multiple center points in the basic design when all other points of the basic CCD have been replicated. Instead, they retain only a single center point in the basic CCD that would be replicated along with all of the other experimental conditions. This plan was followed in the Mills-Williges study and in some of the Williges-North analyses (WN328,3). Dropping multiple center points from a totally replicated CCD was the only true modification of the basic design that Clark and Williges proposed. The result of this change is to degrade the effectiveness of RSM without enough advantage to justify the change. The pros and cons of replicating an entire basic CCD will be discussed later in this paper; here the discussion is concerned only with the consequences when the basic design (replicated or not) fails to include multiple center points.

The number of center points in a CCD affect the following design characteristics and functions:

1. The test for presence of quadratic effects in the first-order model. (BH152,3).
2. The estimate of "pure" error variance needed to test the statistical significance of the lack of fit. (BH169,2).
3. The orthogonality of blocked CCDs. (BH176,4).
4. The "rotatability" of the CCDs. (BH168,2).
5. The uniformity of the "information" profile, (BH168,4).
6. The ability to isolate block and trend effects (Simon, 1974, p. 102).

Quadratic effects. If one or more center points are included along with the hypercube portion of a CCD, the difference between

the mean of the center points and the mean of the 2^k points of the hypercube provides estimates of the sum of the quadratic effects and the variance to be used to test for a lack of fit of the linear model. While this test might be made from data taken at a single center point, data from multiple center points (by each subject) will provide a more stable estimate.

Error estimate. Without overall replication, multiple center points in the basic CCD provide the only estimate of experimental error. This estimate should be made up of the "chance" variability that occurs when the same point is measured several times under the same conditions; it can be contaminated from variability associated with effects that occur when data is tested sequentially.[4] However, when every point in the basic design is replicated, Clark and Williges propose that only a single center point be used in the basic CCD since another source for estimating experimental error would be available (CW306,3). What they fail to indicate is that it would not be an equivalent "experimental error", nor would it be as "pure" an estimate of error.

When there are five variables, as in the Mills-Williges experiment, the basic CCD design would be made up of 30 experimental conditions of which four would have been repeated measures at the center point. In that design, the Subject-by-Center Points variance,

---

[4]Further contamination would occur if an experimenter tested a different subject on each condition (including each repeated center point) of the design. Considering how variable subjects often are, this confounding of subject and conditions differences would ordinarily not be warranted if only a single replication of the design were used.

with 3 x 3 = 9 degrees of freedom, could have been estimated and used as a relatively "pure" estimate of error. However with the replication, Mills and Williges decided to eliminate three of the four center points leaving only 27 conditions in the basic CCD. Therefore, instead of an error term involving the variance of the repeated center points, they used for their error variance a term labelled "Replication" (MW343;343), which was actually the Subjects-by-Experimental Conditions interactions.

Subject-by-Conditions interactions may occur, not by chance, but because such effects often actually exist. They may also occur when truncated, "ceiling and floor" effects are present, and when there are uncontrolled and unisolated sequence effects (trial-to-trial transfer as well as long term trend), and when uncontrolled incidents occur during the data collection. Any argument regarding the purity of this error estimate might have been stronger had linear, quadrantic, and cube trend effects (a total of 9 degrees of freedom) been isolated from the "Replication" term, or had it been demonstrated that the Subject-by-Linear Terms and Subject-by-Quadratic Terms interactions (a total of 25 and 45 degrees of freedom respectively) were not significantly greater than the Subject-by-Lack of Fit term, the most likely term to represent "error." In any case, had the design with multiple center points been used, this entire question of an appropriate error term would have been avoided.

Orthogonality. For orthogonality between estimates of the first and second order coefficients, a certain relationship must exist between the number of center points, the number of experimental conditions in the first and second order blocks, and the value of α

(i.e., the distance from the center to the points of the star) (BH176,4; CW301,2). It is possible to obtain this relationship with only a single center point in the basic CCD design, even if the single center point were located in the star block. But ordinarily, if the investigator intends to use orthogonal blocking along with the iterative approach proposed for RSM, he would begin with the points of the cube block, since the resulting data allow an immediate test of the presence of cross-product, second order effects. Then with center points added to test for possible quadratic effects, he is forced to use multiple center points in his completed CCD since at least one other will be required in the star block and additional ones in the cube portion if it is subblocked. The use of the single center point might be acceptable only if the investigator decided to take the less efficient approach of starting his experiment with the star block first. This entire consideration was avoided, however, in the Mills-Williges study which used neither blocking nor the iterative approach.

Rotatability. A rotatable design is one in which the precision of an estimate is the same at all points equidistant from the center of the experimental space. Rotatability is a primary feature in many of the response surface designs. With CCDs, rotatability is obtained by selecting the proper value for the length of the axis arms of the star, α (BH171,1; CW229,3). However, with exceptions, the α values appropriate for orthogonality and for rotatability, while reasonably close when multiple center points are used, are not equal.

In general, it is agreed that when a decision must be made, the quality of orthogonality is more important to preserve than that of rotatability (BH177,3) (CW301,5). When only a single center point is used for orthogonal blocking, however, the discrepancy between the α for orthogonality and for rotatability increases; if the one for orthogonality is chosen, the rotatable characteristic is further distorted. While not necessarily a serious matter, it is still another degradation that occurs when single center points are included in the basic design.

Uniform information profile. With only a single center point in the basic design, "information" at the center of the response surface will be less precise than at points further from the center. "Information" at any point on the response surface is the reciprocal of the variance at that point (BH166,2). Since the center of the experimental space will ordinarily be that portion in which there is the greatest interest, Box proposes that additional (multiple) center points be included in these response surface designs to make the contour of the information profile approximately constant over the central interval between the two levels of the cube portion. Beyone these points, precision is allowed to degrade considerably (BH169). In the Mills-Williges experiment, with only a single center point in the CCD, the precision of performance estimated at the center of their experimental space is poorer than that estimated away from the center.

Isolating block effects. As stated earlier, with only a single center point, an orthogonally blocked design is possible. However, with a single center point in the CCD, no estimate of block effects

is possible. This means that no trend or other effects that might be confounded with blocks can be isolated, the consequence of which is to distort the estimated error variance. This in turn will affect the tests of statistical significance. It would be optimistic to assume that these effects are negligible in most human factors research. It is only prudent to use methods that will isolate them in the event they occur.

Orthogonalizing quadratic coefficients. In the classic CCDs, the estimates of the coefficients of the quadratic effects are not orthogonal to one another (BH163,1). While Myers (1971, pp. 133-134) describes a way to adjust the design so that this correlation would be eliminated, the adjustment will affect other characteristics of the design and is ordinarily not justified.[5] While not a serious matter when one does not evaluate each term of the equation, the degree of correlation among quadratic terms is greater when only a single rather than multiple center points are used (with other parameters properly adjusted), as shown in Table 4. The consequence of this correlation is to make the estimates of the coefficients of the quadratic terms differ depending upon the particular order in which each is isolated in the analysis.

---

[5]Since this report was prepared, Williges published another paper, "Research Note: Modified Orthogonal Central-Composite Designs", in Human Factors, 1976, 18, 95-97. In this paper he cites Myers' (1971, p. 134) calculations for the alphas required for a completely orthogonal design. However, he failed to note Myers' comment regarding this design, namely: "As we implied previously in this section, there are important choices of α to consider, other than the value which makes the design orthogonal. In many cases, these other choices are more desirable than the orthogonal CCD."

TABLE 4. CORRELATIONS AMONG QUADRATIC TERMS AS A FUNCTION OF NUMBER OF VARIABLES AND NUMBER OF CENTER POINTS IN CCD

Number of Variables in CCD	Number of Center Points Single	Multiple		
Three	-.381	-.090	(6)	Number of Center Points in Unblocked Design
Four	-.282	-.088	(7)	
Five*	-.200	-.067	(6)	
Six*	-.178	-.056	(9)	
Seven*	-.164	-.044	(14)	

*1/2 fraction in cube portion

Deemphasizing Individual Coefficients

It is generally good practice to draw as much information from the experimental results as possible. However, the type of equations generated by response surface designs, such as the CCD, were never intended to be examined term by term. Box and Hunter write:

"Now the primary object of the experimental designs described in this paper is to estimate an unknown response function by means of a mathematical model obtained by using a Taylor's Series expansion of some order. Using such an experimental design, observations are recorded at N points in the factor space, and this evidence is used to estimate the coefficients of the model by least squares. The interest therefore is really directed at the complete estimation equation and not an investigation of the individual estimated coefficients and their variances." (BH165,2).

In CCDs and other response surface designs, the precision of the various estimated coefficients is not constant and, as has already been noted, some coefficients may be correlated. The effects of this are discussed quite thoroughly by Box and Hunter (BH163-167) and are of little concern if the important consideration is the fit of the overall equation. Significance tests are to be applied to pooled estimates of the different orders of the model, i.e., first, second, higher (lack of fit), rather than each individual term. In this regard, Box and Hunter write:

"It should be noted here that the individual coefficients of the model have not been separately tested for significant departure from zero. If this had been done, and one coefficient was found not to be significantly different from zero, we would not be entitled to

replace the given estimate with a zero, for regardless of its magnitude, it is still the best estimate of the unknown coefficient. To replace this estimate by a zero would in effect be replacing a best estimate by a biased one. The important test concerns the order of the model; i.e., whether a model of first order, or of second order, adequately represents the unknown function." (BH174,2).

All of the Williges studies continue to reflect the F-test orientation by examining the statistical significance of each term of the polynomial. Because in two of the studies (WB and WN) only a partial analysis of the collected data was carried out (a fault that was discussed earlier in this paper), the examination of only the linear terms might provide an erroneous interpretation of the reliability of the individual variables. The proper test of the variables, rather than the terms, should have included the unanalyzed second order components. Box and Hunter write the following concerning this procedure:

"Another test that could be run would be to determine whether a particular variable x_i contributed significantly to the response. In this case the sums of squares of all the coefficients bearing an i subscript would be pooled and then tested. However, the search for the important or significant variables should properly preceed [sic] the estimate of a response function by a second order model." (BH174,2).

Contrary to the examples provided in the Williges papers, the last sentence in the above quote emphasizes the strategy whereby the search for important or significant variables should properly

precede the collection of data for fitting a (second order) function. In practice, since the number of candidate variables that conceivably might have a critical effect on performance can be quite large--15 to 30--in most human performance tasks, considerable screening should have taken place prior to the effort to estimate a response surface.

The task of identifying critical variables and the task of relating them functionally should properly be done in two distinct steps; this is the only economical and efficient means of handling truly large numbers of variables (Simon, 1973). The first-order phase of a CCD can be used for the identification purpose, as Box and Hunter suggest, but for truly multifactor research, a more intensive, preliminary screening effort might more practically be carried out.

In the Williges papers, while examining individual terms, the authors fail to warn the reader of the correlation among the quadratic terms. Since the effects of these terms depend on the order in which they are isolated in the regression analysis, the reader should at least realize that any test of significance will be affected to some degree however small. In the Williges-North paper, when analyzing the uncollapsed design, the authors throw out the data collected on each subject for three trials at the center (WN328,3). Since the remaining data had been collected with appropriate α values for the complete design, the analysis is no longer being made on a properly blocked design and estimates of some first and second order coefficients will be correlated.

Interpretation of individual terms under these circumstances is tenuous even if the experimenter is aware of what he has done. Furthermore, the papers do not make it clear that although a single error term might be adequate to test the reliability of the entire equation, its use to test individual coefficients that differ in precision may make interpretation of such an analysis ambiguous. While an examination of results in depth is always desirable, the investigator should be aware of what he is doing and its weaknesses. These are not brought out in the examples in the Williges series.

Finally, Williges and North suggest that one might keep certain "marginally reliable" coefficients if one were searching the experimental space (WN334,1) but not if one wanted the more valid and stable overall prediction equation (WN333,3). As Box and Hunter note, the equation would be biased if marginally significant terms were omitted. It is difficult to understand why a biased equation is more acceptable for purposes of prediction than for search as Williges and North suggest. In the Williges-North paper, the idea of dropping nonsignificant terms is promoted on the grounds of parsimony. But which terms are significant changes in these papers each time more replications are added and would continue to do so until every term would eventually become significant (Bakan, 1966, p. 426; Hays, 1966, p. 326; Kleiter, 1969, p. 10), so it is difficult to know at what point in the program one should decide to drop a term. Under ordinary circumstances, Box and

Hunter's approach of keeping the terms once they have been isolated seems the more manageable and accurate approach for response surface studies.[6]

---

[6]After this report had been prepared, a paper by David J. Cochran and LaVerne L. Hoag, "Response Surface Methodology and Optimization -- A Possible Pitfall," was discovered in the Proceedings of the Human Factors Society 19th Annual Meeting, October 1975. In following the recommendations in the Williges series, these investigators became aware of what they refer to as a "dilemma for which the experimenter is given no method of resolving," namely, the problems of interpretation that arise when "statistically non-significant" terms are dropped from the regression model. Hopefully the discussion in this paper will help them resolve their "dilemma" which was not created by RSM but by following unwise procedures and by the ambiguities inherent in the significance test.

NON-RSM METHODOLOGICAL CONSIDERATIONS

The second major criticism made of this series of papers is that the authors employed poor methodologies not specific to RSM. In some cases this was more or less the result of careless planning; in other cases, however, it occurs as a result of calculated decisions. These cases will be described in detail below.

After summarizing the features of CCDs as developed by Box and his associates, Clark and Williges introduced what they refer to as "modifications" of the basic, blocked, central-composite design (CW295). The modifications are presented as a series of alternatives, the relative advantages of which are determined empirically by the four experiments in the series. Thus they consider the relative advantages of:

> CCDs with multiple observations at only the center point versus CCDs with multiple observations at each experimental point.

Regarding designs of the latter type, they compare the relative value of:

> Analyzing all of the collected data without modification versus collapsing across subjects at each data point prior to analysis and also the relative values of using:
>
> > Between-subjects designs in which no subject is observed more than once and observations at each experimental point might be multiple and unequal or multiple and equal; versus within-subject designs in which each subject is observed only once at each experimental point.

Contrary to what the authors imply, these variations per se do not modify the basic CCD and can be discussed and considered more or less independently of RSM. They are instead alternative procedures that might be used with any basic experimental data collection plan, be it CCD or factorial or lattice square and so forth. In each of these, the methodological considerations are essentially the same. Furthermore, although attempted in this Williges series, the consequences of the alternatives cannot properly be determined empirically, but only through a rational determination based on a knowledge of their statistical and mathematical characteristics. Let us examine each of these alternatives in turn.

Center Point Versus Total Design Replication

Clark and Williges proposed that rather than replicate only at the center of a CCD, every point of the basic design be replicated (CW304,1). Based on an experiment by Williges and Baron, they conclude that total-design replication is better. It will be shown, however, that their implementation of total design replication was neither in accordance with good RSM nor the most economical method of meeting the desired objectives, and that the empirical study actually offered little support for their conclusions regarding this issue.

An investigator may decide to replicate a basic experimental design for either or both of two reasons: to measure performance more precisely and/or to obtain an estimate of experimental error. The former will lead to improved estimates of the coefficients in a regression equation and ultimately the estimates of responses

derived from the equation. The latter may be used to establish confidence limits and to perform tests of statistical significance. Psychologists in general have tended to overuse and misuse replication (Simon, 1973, pp. 19-31), often trading precious time and money replicating rather than studing an expanded experimental space. Many times the replication has been unnecessary and often there are more economical, alternative methods available to meet the desired goals. These criticisms become increasingly pertinent as the number of factors in the experiment increases.

Although the "goodness" of an ANOVA design is partially determined by how well it reduces variable error, discussions of response surface designs have tended to play down concern with variable error. This has been so for two reasons. One reason, as discussed earlier, is that response surface designs also emphasize the reduction of bias error (through improving the fit of the model to the response) on the grounds that a design that is sensitive to bias errors is actually sensitive to both bias and variable error. In this regard, Myers (1971), p. 201) writes: "In fact, it would seem that errors that occur due to bias play an even more important role, as far as [the estimated response] $\hat{y}$ is concerned, than those errors which result from sampling variation." Earlier he had noted that only when the variable contribution is more than six times the bias would an experimental design, totally concerned with bias error, not be adequate.

Discussion of variable error has also been minimized in many papers on RSM because these techniques were first applied in chemical rather than agricultural or human performance studies.

In the former, responses tend to be more reliable than in the latter types making variable error less of a problem. However, Box and Hunter do not totally ignore the issue for they write in accordance with good RSM principles: "In some examples the large size of the experimental error would make it essential to replicate the experiments. If the size of the experimental error is not known it is best to proceed sequentially, performing further experiments if the standard errors of the coefficients estimated from the first set are too large" (BH144, footnote).

Thus, unlike the Williges studies in which the decision to make multiple replications of the basic design preceded any data collection, in RSM methodology each replication is considered a new experiment to be added only after examination of the previously collected data suggests that it is warranted. As we shall see, even when the need for some replication can be anticipated, the massive replication approach proposed by Clark and Williges and used in the illustrative studies is not the most economical.

But in the above quote, Box and Hunter were concerned only that the general magnitude of the experimental error of the observed responses might be large and should be reduced with replication throughout the design. Clark and Williges properly point out the possibility that the experimental error of the observed responses might be unequal in different parts of the experimental design. They write: "When the goal is to approximate an entire response surface (rather than merely that portion of the surface surrounding the optimum), limiting multiple observations to a single experimental point may not be the most judicious

strategy. Indeed, the actual variability in response may be so great across subjects and data points that it would be unrealistic to presume the standard of estimate at the center point is an adequate estimate of error at all points" (CW304,1).

However, except for repeating essentially the same comment later in their paper (CW305,3), Clark and Williges never again consider the problem of heterogeneity of variance of the observed responses.[7] Nor is there any discussion of the issue nor how it was handled in any of the experimental studies in that series which used total design replication to offset this potential effect. If in these studies the variance of the observed responses did in fact differ at different parts of the experimental design (as Clark and Williges suggest might happen), then it was no more proper to use the error estimate from these composite but heterogeneous variances than it would have been to use the estimate based only on the replicated center points. Neither estimate would have been representative nor suitable for performing a test of significance.

---

[7]Clark and Williges do not make it clear to their readers that, in CCDs, even if the variance at every observation point of the experimental design were essentially equal, neither the variances of the beta coefficients in the regression equation nor the variance of the estimated responses throughout the response surface would be equal. Box and Hunter were not concerned with the relative precision of the estimated beta coefficients of the second order model -- they are not equally precise -- for they consider this to be "the wrong question" (BH163,2). Nor are unequal variances at different points across the response surface an issue since rotatable designs only require that points equidistant from the center have equal variance. Variability increases considerably in correctly designed CCDs beyond the ± 1(coded) points in the design (BH169). Of course, even in classical factorial designs the precision varies markedly across the response surface (BH166,2).

When Box and Hunter proposed using the replicated center point for estimating error variance to test the lack of fit they did so with "the usual assumption that the variances of all determinations are equal" (BH169,2). When this assumption is not met, it is not correct to combine the heterogeneous variances. One advantage of the iterative approach of RSM is that this heterogeneity would be discovered early enough to permit some scale transformations to be introduced to correct the matter before an expensive, massive replication had taken place.

Now on the other hand, if in the Williges studies the observed variances were found to be homogeneous after all, the failure to use the iterative RSM approach to replication (as well as to model building) could cost a great deal in wasted effort. Even a few preliminary tests at selected points in the design might have been a more economical way to determine the need to be concerned with both the magnitude and the heterogeneity in performance variability.

Clark and Williges write that the Williges-Baron study "affords a striking demonstration of the effect of estimating experimental error at a single replicated point as opposed to estimating it across a series of replicated points" (CW304,1). Actually, the study did not consider the original issue of variance heterogeneity at different points in the experimental design. Instead, what this empirical effort "demonstrated" was that "when replications were restricted to the center points, none of the experimental factors was found to contribute significantly to the response level, despite their apparent importance in the resulting prediction

equation. When multiple observations were made at each of the data points, however, the subsequent analysis revealed that some of the experimental variables were significant in determining the response level."

These statements are true in fact but false in implication. All this study demonstrated was the obvious fact that when the degrees of freedom for the error term are increased, the significance test becomes more sensitive. Making the point that many have made, Hays (1963, p. 326) states: "Virtually any study can be made to show significant results if one uses enough subjects, regardless of how nonsensical the content may be.... This kind of testmanship...clutters up the literature with findings that are often not worth pursuing, and which serve only to obscure the really important predictive relations that occasionally appear."

Nunnally (1960, p. 643) reiterates the same point by saying: "If the null hypothesis is not rejected, it is usually because the N is too small. If enough data are gathered, the hypothesis will generally be rejected." Certainly no empirical effort is required to illustrate this fact, and, more to the point at hand, it does not decisively demonstrate the relative merits of the two procedures since the particular effect that Clark and Williges use as proof for their conclusion also could have been achieved by replicating the center point (in this example) twenty more times.

When the variability of the observed responses is suspected of being larger than desirable and the possibility of variance heterogeneity throughout the design is a concern, Dykstra (1960) proposes using partially duplicating response surface designs.

Combined with the iterative approach of RSM, these plans provide essentially the same information that the Clark-Williges massive replication plan offered and do so far more economically. When truly multifactor experiments are conducted, this saving can become considerable.

Analyzing Collapsed Versus Uncollapsed Data

The major purpose of the Williges-North paper, they say, is "methodological" (WN323,3). Clark and Williges (1973) discussed two ways of analyzing data collected from a completely replicated RSM central-composite design. One, all of the data could be analyzed directly, or alternatively, the data could be collapsed across subjects prior to analysis, thereby reducing the design to the equivalent form of an unreplicated, basic RSM central-composite design with repeated observations only at the center. These alternate analyses were compared in the Williges-North study in terms of their resulting sensitivity and in terms of the predictive validity of the regression equation as determined through cross-validation.

Two conclusions cited by Williges and North were that the uncollapsed designs produced a more sensitive F-test than collapsed designs and that uncollapsed designs gave more realistic predictions than collapsed designs (WN334,3). In the discussion that follows it will be shown that the first conclusion is inherent in the F-test and needs no empirical verification and that the second conclusion is not supported by the data.

Design sensitivity arguments. As the investigators themselves noted (WN329,3), the analysis with the uncollapsed data was more

sensitive--which meant that more terms were found to be statistically significant--than the collapsed data because of the drop in degrees of freedom--from 120 to 3--in the error term after collapsing. Just why the investigators felt the need to perform an empirical study to demonstrate this fact is unclear. In the preceding Williges-Baron study they had discovered (?) that total replication had increased the degrees of freedom in the error term thus causing a more sensitive F-test.

Now in this Williges-North study, they reverse the procedure --since averaging across subjects is essentially equivalent to removing replication--and lose degrees of freedom in the error term and consequently sensitivity in the F-test. Later, when the results from the cross-validation studies are combined with the original data, i.e., essentially adding still more replications to the uncollapsed data, the F-test becomes even more sensitive (WN331,5). Since the value required for a significant F decreases as the number of degrees of freedom in the error term increases, these results could have been predicted without any empirical study. Insofar as that conclusion is concerned, the experiment was irrelevant.

Of a more serious concern, however, is the interpretation implied by the investigators in both studies (WB and WN), namely that the design that obtains the most statistically significant terms is necessarily the better one. But it is not a suitable criterion; in fact, as Lykken (1968, p. 158) says: "...statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition for concluding

that... a useful empirical fact has been established..." Hays (1963, p. 300) states: "It is a grave error to evaluate the 'goodness' of an experiment only in terms of the significance levels of its results... it is entirely possible for a highly significant result to contribute nothing to our ability to predict behavior, and for a nonsignificant result to mask an important gain in predictive ability."

Dunnette (1966, p. 345) comments how most psychologists "still remain content to build our theoretical castles on the quicksand of merely rejecting the null hypothesis" and Nunnally (1960, p. 650) warns: "We should not feel proud when we see the psychologist smile and say 'the correlation is significant beyond the .01 level'. Perhaps that is the most he can say, but he has no reason to smile." Campbell and Stanley (1963, p. 22) sum it up quite simply by saying: "Good experimental design is separable from the use of statistical tests of significance."

In the context of CCDs, the primary purpose of the significance test is to discover the adequacy of fit of the equation, and even for this purpose, as stated earlier, it is best used merely as an adjunct clue after examining the relative proportions of the performance variance accounted for by the regression and by what's left over.

Cross-validation arguments. Williges and North performed cross-validation studies and concluded that "...uncollapsed or within-subject analyses as suggested by Clark and Williges (1973) appear to provide a more sensitive analysis as well as more realistic estimates of the predictive worth of the regression

equations as compared to collapsed analyses when predictions of individual performance are made" (WN334,3). A more straightforward conclusion appears in their summary, namely, "the uncollapsed, within-subject designs provided the better prediction equations" (WN321, summary) as compared to collapsed designs. The point of discussion here will not be whether one form of the equation or the other is in fact better but whether the investigators properly interpreted their data and whether they employed the methodology that would permit this type of conclusion to be drawn at all.

The basis for the conclusion drawn by Williges and North was not, as is usually the case, how well the equations, derived from one data sample, predicted performance obtained from a second data sample. Instead, it was how well the correlations between predicted and observed performance from two samples agreed with estimated population correlations derived by applying a "shrinkage" formula to the data from the first sample. The greater the difference between the empirical and theoretical correlations, with the latter being used as the standard of goodness, the poorer Williges and North concluded their empirical results to be. Essentially what these investigators seem to be claiming is that equations that ought to do better (but didn't) are better than equations that did do better (but oughtn't to have according to a formula of questionable merit). There are several formulae for estimating shrinkage, each with its own assumptions and limitations.

Although they had originally used empirical results to evaluate and select their shrinkage formula (North and Williges, 1972, p. 221), they now use the theoretical estimations to evaluate

the empirical. Shrinkage formulae are intended to be used in lieu of further empirical tests. Although many questions can be raised concerning the usefulness of cross-validation studies (Smith, 1970), nevertheless, if a competent data collection program has been undertaken, then the fact rather than the theory should be the criterion by which the equations (and designs) are to be evaluated. The question is simply: which analysis--of uncollapsed or collapsed data--produced the equations that predicted the actual performance from a second set of data more accurately?

In Table 5, representative data extracted from Tables 5 and 6 in the Williges-North paper (WN333 and 334) are presented for one condition, "Latency response with the black-and-white TV system." There seems to be little question that equations derived from the collapsed data in general always estimated observed performance as well or better than equations derived from uncollapsed data. This is true both within and between samples. This should not come as any surprise since in the collapsed data, one major source of uncontrolled variability--subject differences--has been removed. Yet Williges and North concluded otherwise.

But whether or not the Williges-North data had been properly interpreted was actually a moot point. The data collection methodology in this paper was so confused that any conclusions regarding the relative merits of collapsed versus uncollapsed data would be questionable because of the other conditions irreconcilably confounded with these two alternatives. The following are the more obvious examples:

TABLE 5. CORRELATIONS BETWEEN OBSERVED PERFORMANCE DATA AND VALUES ESTIMATED FROM EQUATIONS BASED ON DIFFERENT REGRESSION MODELS*

Regression Model of Estimation Equation Derived From First Data Sample	Source of Observed Performance Data		
	Results from First Sample	Results from Second Sample	
		Collapsed	Uncollapsed
Collapsed, 2nd order model	.870	.687	.438
Collapsed, 1st order model	.779	.688	.433
Uncollapsed, 2nd order model	.561	.438	.425
Uncollapsed, 1st order model	.464	.450	.450

*This data was taken from Tables 5 and 6 of the Williges-North (1973) paper. It is only the data for the Latency response scores for the black and white TV system, yet it is quite representative of all the data.

1. The collapsed equations are based on median performance measures while the uncollapsed equations are based on mean performance measures. If the subject data is skewed and/or skewed differently for different experimental conditions, then the equations could predict differently without regard for the collapsing issue per se.
2. The designated "error" variance for the collapsed equation was actually an average within-subjects variability of measures all taken at the center of the experimental space. For the uncollapsed equations, the designated "error" variance was actually the interaction between subjects and the entire set of experimental conditions. Using different definitions of "unexplained" variance affects the proportion of total variance accounted for by the equations and differentially affects the tests of statistical significance based on this error variance.
3. All data initially collected were included in the derivation of the collapsed equation, while sixteen percent of the data were excluded from the derivation of the uncollapsed equation. The excluded data had come from the center points the investigators judged were superficial.
4. In the collapsed equation, the coefficients of all first and second order terms were independent of block effects and the effects of one another. In the uncollapsed equations, first and second order terms were biased to some degree since dropping the center points destroyed the orthogonality of the CCD being used.

5. The investigators, concerned with possible "sequence" effects, stated that they counterbalanced the order in which the blocks were administered although any mean differences among blocks would have been neutralized anyway with the orthogonally blocked CCD. On the other hand, they did not indicate the method used to control unwanted sequence effects which are likely to occur when a subject is tested serially on the ten conditions within blocks. Since complete counterbalancing of the serial order of ten conditions with only six subjects, as used by Williges and North, is not possible, any sequence effects that may have occurred would differentially affect the two equations. One source of sequence effects is confounded with the subject-by-conditions interaction and, if not properly isolated, would distort the main effects of both sets of data and inflate the error term in the uncollapsed data.

Since none of the above is an inherent characteristic of collapsing or not collapsing data, the confounding of conditions prevents clear-cut assessment of the relative merits of these two methods of analysis from the data presented.

Between-Subject vs Within-Subject Designs

Clark and Williges state that "when noncollapsed designs are used, the investigator must make another major design decision with respect to his selected design. If, due to the nature of his research problem, he chooses to observe different subjects at each

of the experimental points, the resulting study constitutes a between-subjects design. If, on the other hand, he elects to observe each subject under all experimental conditions, the resulting study constitutes a within-subject design. The choice of a between versus a within-subject design is dictated by the particular question which the researcher is investigating. In either case, if the necessary restrictions are observed, the design conforms to the basic central-composite design" (CW305,3).

Of course, whether the same or different subjects are used is a methodological question that is independent of RSM and CCDs, and that could be made not only "when noncollapsed designs are used" but also when collapsed designs are used. Furthermore, there is a third alternative available to an experimenter concerned with the serial assignment of experimental conditions to subjects, which has certain methodological advantages not mentioned in the Williges series. Thus, different groups (as well as numbers) of subjects may be used in each block and under the proper conditions could be used not merely as a means of building up the degrees of freedom of the error term, but to control and isolate sequence effects within blocks.

This experimental strategy was illustrated in a study by Mueller and Simon which is described in a paper by Simon (1970b). Although Clark and Williges (CW307,3) warn of the importance of "proper counterbalancing" in within-subjects designs "so as to avoid spurious sequence effects," except between blocks where it should not matter when correctly orthogonalized designs are em-

ployed, proper counterbalancing within blocks was neither described nor employed in the two papers of the series (WN and MW) using within-subject designs.

EVALUATING RESPONSE SURFACE METHODOLOGY

The third major criticism of the series was that the authors offered no evaluation of the CCD in the context of RSM. Throughout the series, it is implied that in addition to illustrating RSM and testing certain variations to the CCD, the studies also represent an empirical evaluation of the usefulness of these tools in human performance experiments. Thus comments such as the following quotations are found in the conclusions or the summaries of the papers in the series:

"The results of this study clearly indicate that RSM techniques provide both a useful and economic approach for investigating the effects of several variables on human transfer performance." (WB318,3).

"It is clear from the results that RSM central-composite design techniques are successful in providing efficient procedures for generating multiple-regression prediction equations for variables important in cartographic symbol locations tasks." (WN335,2).

"The utility of this approach was demonstrated in that it provided efficient data collection, and the observations obtained from the response surface equation described complex relationships among the five parameters investigated." (MW348,2).

"An RSM central composite design provided an efficient method for obtaining data and quantifying the relationship." (WM349, summary).

In fact, none of the investigations was designed in a way that could experimentally evaluate CCDs in the context of RSM.

Two studies in the series were oriented particularly to the evaluation role. Thus, Williges and Mills (WM349,3) stated that the purpose of their study "was to investigate the predictive validity of the RSM regression equation" from a different point of view than had been employed for the same purpose by the Williges-North study. Williges and Mills determined how well the estimates from an equation derived from one set of data correlated with observed performance values obtained from the same subjects at new points in the same experimental space. In the Williges-North study, after the initial data collection effort, a second set of data was collected at the same coordinates in the experimental space but with different subjects. They determined how well the estimates from the equations derived from the original data correlated with performance obtained in the second effort.

Now the procedure in both studies was essentially to collect data from sample data points within the experimental space, derive a multiple regression equation based on those data, and then see if that equation could estimate a second set of data taken at the

same or equivalent points in the same space.[8] To reduce this situation to its least common denominator, imagine that instead of the points of a CCD, only a single data point had been treated to the above procedure. Obviously then the retest effort is merely a measure of reliability (when we make an untested assumption that the two sets of subjects are homogeneous). The same is true when retesting is done with the larger number of points of a CCD or any other design. It is only the reliability of the data that is being measured along with the experimenter's ability to eliminate measurement and sampling errors and to control for unwanted effects that might occur when the data are being collected.

There is no measure of "predictive validity" nor of the effectiveness of the CCD. Since there was no effort to compare performance estimates from the equation with performance under real world operational conditions, no test of the predictive validity of the equations was made. Since no other configuration of

---

[8] Williges and Mills (MW), for their "cross-validation" test, collect the second set of data from the other half of the 2^5 factorial, the first half of which had been used in the cube portion of the original CCD design. They imply that by examining points interpolated among the original set, they are doing a different evaluation than Williges and North had done when they used the same points. But this is not so, if the basic assumption of the CCD is met, namely, that a second order model will adequately fit the data. If a second order equation adequately fits the data, then estimates of all main and two-factor interaction effects, whether estimated from points for one or the other half of the 2^{5-1} (Resolution V) fractional factorial, should be identical within the limits of the reliability of the measurements. This is so by definition. Of course, if that assumption is not met, then the basic principle of RSM -- to continue collecting data to estimate higher order effects until the data is fit -- has not been satisfied. This note does not deny that testing the other half of fractional factorial is preferable over repeating the original half. However there would be no advantage had the experiment satisfied the RSM principle of data fitting as it is supposed to.

experimental data collection points was compared with that of the CCD, no test of the relative effectiveness of CCDs was made. As stated earlier, regression equations can be derived from any set of data. Evaluating experimental designs requires a test that will determine whether sampling the data from the experimental space according to one pattern will result in a more accurate representation of the response surface than sampling the data according to another pattern. There are many other patterns that might be used in lieu of CCD and compared for both economy and efficiency, none of which was ever considered in the Williges series.

Other investigators have compared the CCD with other data collection patterns (Box and Hunter, 1958; Brooks, 1955; DeBaun, 1959). However all employ analytic techniques since an evaluation of this sort cannot properly be made empirically.

EPILOGUE

A paraphrase of a quote from John Gardner (1961) would seem to be an appropriate way to close:

"The society which scorns excellence in plumbing, because plumbing is a humble activity, and tolerates shoddiness in [research] because it is an exalted activity, will have neither good plumbing nor good [research]. Neither its pipes nor its theories will hold water." (p. 86).

REFERENCES

Bakan, D. The test of significance in psychological research. Psychological Bulletin, 1966, 66, 423-437.

Box, G. E. P., and Hunter, J. S. Experimental design for the exploration and exploitation of response surfaces. In Chew, V. (Ed.), Experimental design in industry. New York: John Wiley, 1958, 138-190.

Box, G. E. P., and Wilson, K. P. On the experimental attainment of optimum conditions. Journal of the Royal Statistical Society, Series B, 1951, 13, 1-45.

Brooks, S. Comparison of methods for estimating the optimal factor combination. Sc.D. thesis. John Hopkins University, 1955. Cited by W. G. Cochran and Gertrude M. Cox, Experimental designs. (2nd ed.) New York: John Wiley, 1957, 367-368.

Campbell, D. T., and Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Clark, Christine, and Williges, R. C. Response surface methodology central-composite design modifications for human performance research. Human Factors, 1973, 15, 295-210.

Das, M. N., and Narasimham, V. L. Construction of rotatable designs through balanced incomplete block designs. Annual of Mathematical Statistics, 1962, 33, 1421-1439.

DeBaun, R. M. 1959. Response surface designs for three factors at three levels. Technometrics, 1959, 1, 1-8.

Dunnette, M. D. Fads, fashions, and folderol in psychology. American Psychologist, 1966, 21, 343-352.

Dykstra, Jr., O. Partial replication of response surface designs. Technometrics, 1960, 2, 185-195.

Gardner, J. W. Excellence, can we be equal and excellent too? New York: Harper, 1961.

Hays, W. L. Statistics. New York: Holt, Rinehard, and Winston, 1963.

Hill, W. J., and Hunter, W. G. A review of response surface methodology: a literature survey. Technometrics, 1966, 8, 571-590.

Kleiter, G. Krise der signifikanztests in der psychologie (The crisis of significance tests in psychology). Jahrbuch für Psychologie, Psychotherapie und Medizinische Anthropologie, 1969, 17, 144-163. Translated by D. P. Barrett, Royal Aircraft Establishment Library Translation 1649, The R.A.E. Library, Q.4 Bldg., R.A.E. Farnborough Hants, England, June 1972.

Lykken, D. T. Statistical significance in psychological research. Psychological Bulletin, 1968, 70, 151-159.

Meyers, D. L. Response surface methodology in education and psychology. Journal of Experimental Education, 1963, 31, 330-336.

Mills, R. G., and Williges, R. C. Performance prediction in a single-operator simulated surveillance system. Human Factors, 1973, 15, 337-348.

Myers, R. M. Response surface methodology. Boston: Allyn and Bacon, 1971.

North, R. A., and Williges, R. C. Video cartographic image interpretability assessed by response surface methodology. Savoy, Ill.: University of Illinois, Aviation Research Laboratory Technical Report ARL-71-22/AFOSR-71-8, October 1971.

North, R. A., and Williges, R. C. Double cross-validation of video cartographic symbol location performance. Proceedings of the Sixteenth Annual Meeting of the Human Factors Society, 1972, 220-230.

Nunnally, J. The place of statistics in psychology. Education and Psychological Measurements, 1960, 20, 641-650.

Simon, C. W. Reducing irrelevant variance through the use of blocked experimental design. Culver City, Calif.: Hughes Aircraft Company, Technical Report AFOSR-70-5, November 1970a. 65 pp. (AD 776-041).

Simon, C. W. The use of central-composite designs in human factors engineering experiments. Culver City, Calif.: Hughes Aircraft Company, Technical Report AFOSR-70-6, December 1970b. 52 pp. (AD 748-277).

Simon, C. W. Economical multifactor designs for human factors engineering experiment. Culver City, Calif.: Hughes Aircraft Company, Technical Report P73-326, June 1973, 171 pp. (AD 767-739).

Simon, C. W. Methods for handling sequence effects in human factors engineering experiments. Culver City, Calif.: Hughes Aircraft Company, Technical Report P74-451, December 1974. 197 pp. (AD A006-240).

Simon, C. W. Methods for improving information from "undesigned" human factors experiments. Culver City, Calif.: Hughes Aircraft Company, Technical Report P75-287, July 1975. 82 pp. (AD A018-455).

Smith, Jr., N. C. Replication studies: a neglected aspect of psychological research. American Psychologist, 170, 25, 970-975.

Williges, R. C., and Baron, M. L. Transfer assessment using a between-subjects central-composite design. Human Factors, 1973, 15, 311-319.

Williges, R. C., and Mills, R. G. Predictive validity of central-composite design regression equations. Human Factors, 1973, 15, 349-354.

Williges, R. C., and North. Prediction and cross-validation of video cartographic symbol location performance. Human Factors, 1973, 15, 321-336.

REPORTS PREPARED ON THE "ADVANCED METHODOLOGIES" PROGRAM

Simon, C. W. Reducing irrelevant variance through the use of blocked experimental designs. Hughes Aircraft Company, Technical Report No. AFOSR-70-5, November 1970. 65 pp. (AD 776-041)

Simon, C. W. The use of central-composite designs in human factors engineering experiments. Hughes Aircraft Company, Technical Report No. AFOSR-70-6, December 1970. 52 pp. (AD 748-277)

Simon, C. W. Considerations for the proper design and interpretation of human factors engineering experiments. Hughes Aircraft Company, Technical Report No. P73-325 (Draft No. TR-ARL-71-27/AFOSR-71-11), December 1971. 135 pp.

Simon, C. W. Experiment simulation. Hughes Aircraft Company, Technical Report No. ARL-72-7/AFOSR-72-3, April 1972. 48 pp. (AD 754-215)

Simon, C. W. Economical multifactor designs for human factors engineering experiment. Hughes Aircraft Company, Technical Report No. P73-326A, June 1973. 171 pp. (AD 767-739)

Simon, C. W. Methods for handling sequence effects in human factors engineering experiments. Hughes Aircraft Company, Technical Report No. P74-451A, December 1974. 197 pp. (AD A006-240)

Simon, C. W. Methods for improving information from "undesigned" human factors experiments. Hughes Aircraft Company, Technical Report No. P75-287, July 1975. 82 pp. (AD A018-455)

Simon, C. W. Analysis of human factors engineering experiments: characteristics, results and applications. Canyon Research Group, Inc., Technical Report No. CWS-02-76, August 1976. 104 pp.

Simon, C. W. Response surface methodology revisited: a commentary on research strategy. Canyon Research Group, Inc., Technical Report No. CWS-01-76, July 1976. 60 pp.

DISTRIBUTION LIST

LCdr. James Ashburn, MSC, USN
NAMRL, Bldg. 1953
Pensacola, FL 32512

Dr. L. E. Banderet
SGDR-UE-CR
Dept. of the Army
U. S. Army Research Institute
of Environmental Medicine
Natick, Mass. 01760

Mr. Vernon E. Carter
Pilot Training Systems
Orgn 3750/62
Northrop Corp./Aircraft Div.
3901 W. Broadway
Hawthorne, CA 90250

Dr. Julien M. Christensen
Chairman, Dept. of Industrial Engr.
Wayne State University
Detroit, Michigan 48202

Mr. James Duva (N-215)
Naval Training Equipment Ctr.
Orlando, FL 32813

Dr. Gordon A. Eckstrand
AFHRL/AS
Wright-Patterson AFB OH 45433

Mr. Ronald A. Erickson, Code 3175
Head, HF Branch, Weapons Devel. Dept.
U. S. Naval Weapons Center
China Lake, California 93555

Dr. Marshall J. Farr
ONR, Code 458
800 N. Quincy Street
Arlington, VA 22217

Terrence W. Faulkner
Health & Safety Laboratory
Bldg. 56, Kodak Park Division
Eastman Kodak Co.
Rochester, N.Y. 14650

Mr. Charles A. Gainer
Chier, Army Research Unit
Bldg. 502, P.O. Box 428
Ft. Rucker, Alabama 36360

Dr. Robert A. Goldbeck
Mail Station S-32
Western Development Laboratories
Division
Philco-Ford Corporation
3939 Fabian Way
Palo Alto, California 94303

James E. Goodson, CDR MSC USN
Head, Aerospace Psychology Dept.
Code 15, Naval Aerospace Med. Research Lab.
Pensacola, Florida 32508

Dr. Tom Gray
AFHRL/FT
Williams AFB, AZ 85224

G. C. Helmstadter, Director
University Testing Services
Payne Hall, B302
Arizona State University
Tempe, Arizona 85281

Dr. Charles O. Hopkins
Head, Aviation Research Lab
University of Illinois
Willard Airport
Savoy, Illinois 61874

Dr. Richard Jagacinski
Human Performance Center
330 Packard Road
Ann Arbor, Michigan 48104

Dr. Edgar M. Johnson
U. S. Army Research Institute for the
Behavioral and Social Sciences
1300 Wilson Blvd.
Arlington, VA 22209

AFHRL/ASM (Patricia A. Knoop)
Wright-Patterson AFB OH 45433

Dr. Richard L. Krumm
P. O. Box 2706
Main Post Office
Washington, D.C. 20013

Mr. Robert G. Mills
6570th AMRL/HEB
Wright-Patterson AFB, OH 45433

Distribution List (Continued)

Dr. Frederick A. Muckler (Code 311)
Prog. Dir., Design of Manned Systems
Navy Personnel R. & D. Center
San Diego, CA 92152

Dr. Wallace W. Prophet
Director, HumRRO Cent. Div.
400 Plaza Bldg.
Pensacola, FL 32505

Dr. James J. Regan
Navy Personnel R&D Ctr.
San Diego, CA 92152

Dr. Clyde R. Replogle
6750 AMRL/EME
Wright-Patterson AFB OH 45433

Charles V. Riche
School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332

John E. Robinson, Jr.
Human Factors Staff
Building 606, M.S. G-233
Hughes Aircraft Co.
Fullerton, CA 92634

Dr. Marty Rockway
Technical Director
AFHRL/TT
Lowry AFB, CO 80230

Dr. Stanley N. Roscoe
Bldg. 6, MS D-120
Hughes Aircraft Co.
Culver City, CA 90230

Dr. Mark S. Sanders
Department of Psychology
California State University
Northridge, CA 91324

Dr. Dennis E. Smith
Mathematical Statistician
Desmatics, Inc.
P. O. Box 863
State College, PA 16801

Dr. Margaret J. Smith
Naval Education and Training
Program Development Center
Ellyson Field
Pensacola, FL 32509

H. C. Strasel
Chief, ARI Field Unit
U. S. Army Research Institute
P. O. Box 2086
Ft. Benning, GA 31905

Dr. Martin A. Tolcott
Human Engineering Div., ONR
800 N. Quincy Street
Arlington, VA 22217

Dr. Donald A. Topmiller
AMRL/HES
Wright-Patterson AFB OH 45433

AMRL/HE (Dr. Melvin J. Warrick)
Wright-Patterson AFB, OH 45433

Dr. R. Young, Director
Human Resources Office, ARPA
1400 Wilson Blvd.
Arlington, VA 22209

HQ AFSC/DLS
Andrews AFB, MD 20334

ERIC
Processing and Reference Facility
4833 Rugby Ave., Suite 303
Bethesda, MD 20014

HQ AFHRL/CC
Brooks AFB, TX 78235

Director, Behavioral Sciences Dept.
USAF Academy
Colorado Springs, CO 80840

Flight Dynamics & Control Division
Mail Stop 152
NASA - Langley Research Center
Hampton, VA 23665
Attn: Gary P. Beasley

Department of the Air Force
Air Force Human Relations Lab. (AFSC)
Lackland AFB, TX 78236
Attn: Mark Nataupsky, Capt., USAF
Chief, Evaluation Section
Personnel Research Division

Distribution List (Continued)

Military Asst. For Human Resources
OAD (E&LS)
ODDR&E
Pentagon, Washington, D.C. 20330

Executive Editor
Psychological Abstracts
American Psychological Assn.
1200 17th St. N.W.
Washington, D.C. 20036

HQ USAF/RDPS
Washington, D.C. 20330

AFFDL/CC
Wright-Patterson AFB, OH 45433

Director
USAF Avionics Laboratory
Wright-Patterson AFB, OH 45433

AMD/RDH (Col. George C. Mohr)
Brooks AFB, TX 78235

Dr. Howard L. Parris
AFHRL/CCS
Brooks AFB, Texas 78235

Defense Documentation Center
Cameron Station
Alexandria, VA 22314

Education Research Information Center
Processing & Reference Facility
4833 Rugby Ave., Suite 303
Bethesda, MD 20014

NASA - Scientific & Technical Information Facility
P. O. Box 8757
B.W.I Airport, Maryland 21240

National Technical Information Services (NTIS)
Operations Division
5285 Port Royal Road
Springfield, VA 22151

Maj. Brian Waters
Branch Chief, AFHR K, TTT
Lowry AFB, Colorado 80230

Dr. Robert C. Williges
Department of Industrial Engineering
Virginia Polytechnic Institute and
State University
130 Whittemore Hall
Blacksburg, VA 24061

Dr. Chriss Clark
Honeywell, Inc. (MS R-2340)
2600 Ridgway Parkway
Minneapolis, Minn. 55413

Dr. Robert A. North
Naval Aerospace and Medical
Research Laboratory
Naval Air Station
Pensacola, Florida 32508